The (Technological) Future of Student Surveys

Prof. Dr. Jan Karem Höhne
German Centre for Higher Education Research and Science Studies (DZHW)
Leibniz University Hannover

Keynote: EUROSTUDENT Researchers' Forum 2026
Malta – March 2 to 4, 2026

DZHW

Scientific Director:
Dr. Marcus Beiner

Administrative Director:
Regina Oelfke

Department 1
Educational Careers and Graduate Employment

Acting Head:
Dr. Frauke Peter

Deputy Head:
Dr. Kristina Hausschildt

Department 2
Research System and Science Dynamics

Head:
Prof. Dr. Emanuel Kulcyzcki

Deputy Head:
Dr. Guido Speiser

Department 3
Governance in Higher Education and Science

Head:
Prof. Dr. Bernd Kleimann

Deputy Head:
Frank Dölle

Junior Professorship:
Prof. Dr. Nikolas Eisentraut

Department 4
Research Infrastructure and Methods

Head:
Dr. David Broneske

Deputy Head:
Dr. Karsten Stephan

CS3 Lab:
Prof. Dr. Jan Karem Höhne

Department 5
Administration

Head:
Regina Oelfke

Research Clusters

Research Data Centre (FDZ)

Head:
Daniel Buck

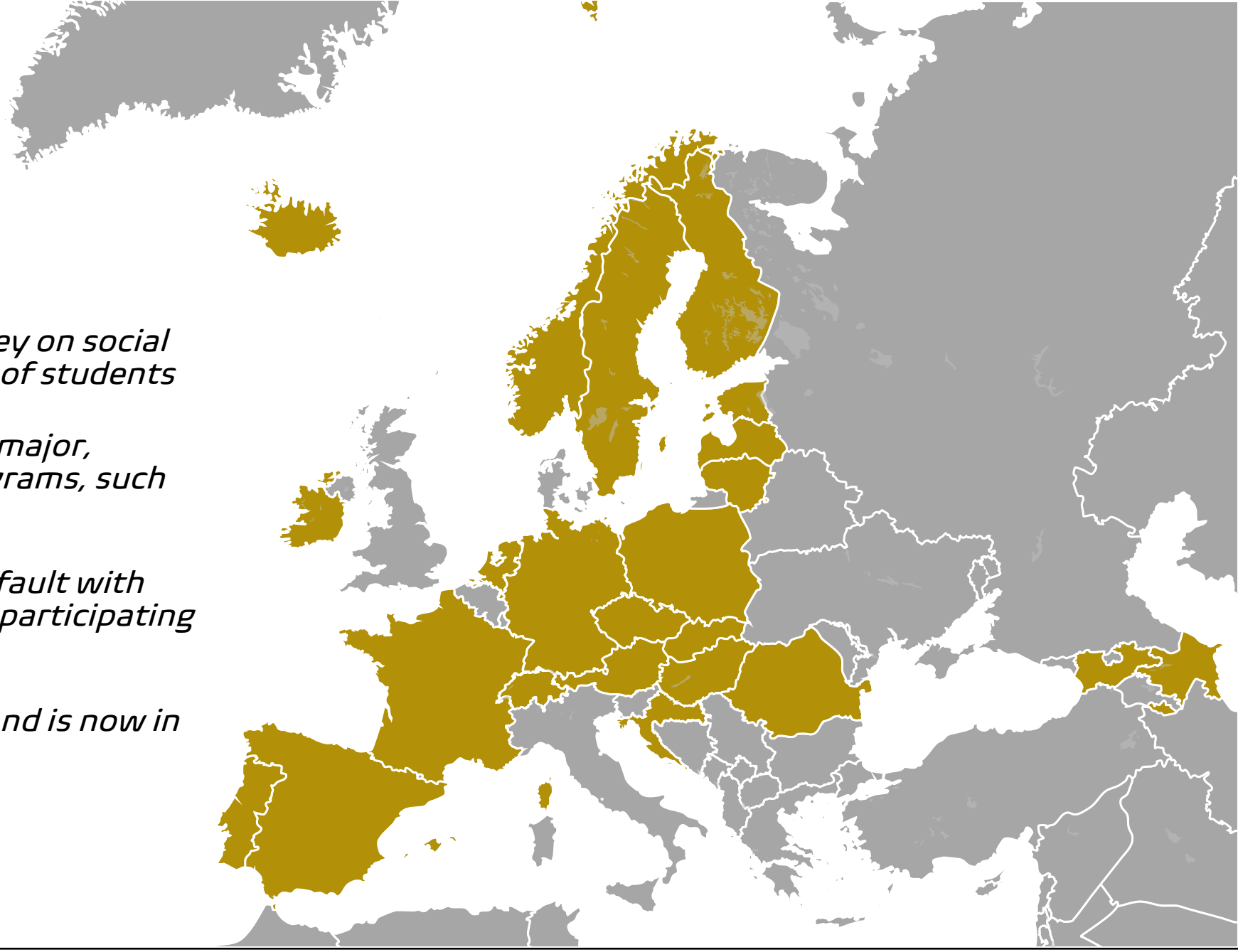
I am grateful to the outstanding members of my
CS3 lab for Computational Survey and Social Science:

Joshua Claassen

Lisa Steiner

Ben Wolf

- EUROSTUDENT …
  - *… is a cross-country survey on social and economic conditions of students*
  - *… is comparable to other major, international survey programs, such as ESS and SHARE*
  - *… uses a web mode by default with national sampling across participating countries*
  - *… runs since 1999/2000 and is now in its 9th round*

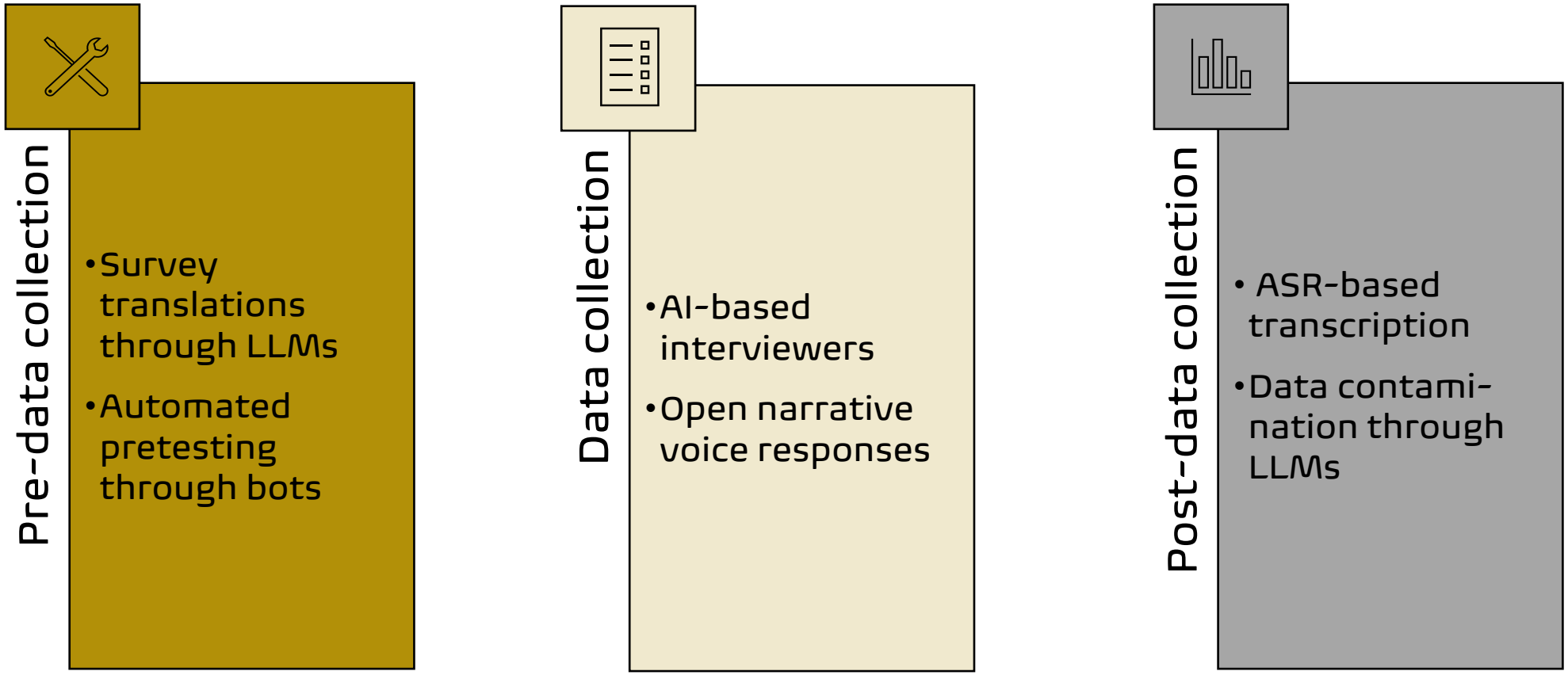


Quagmire of predictions of the future

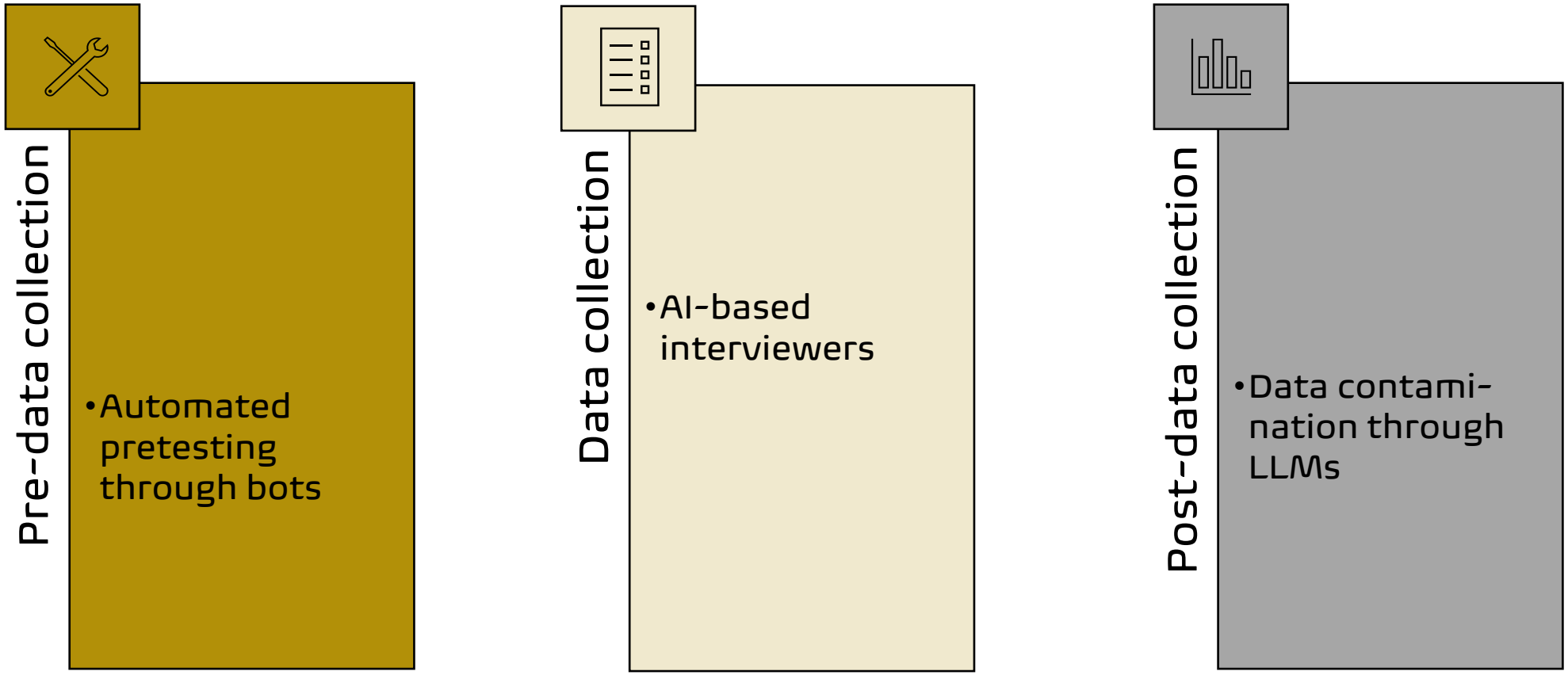
An alleged quote from Bill Gates:

"640 KB ought to be enough for anyone"

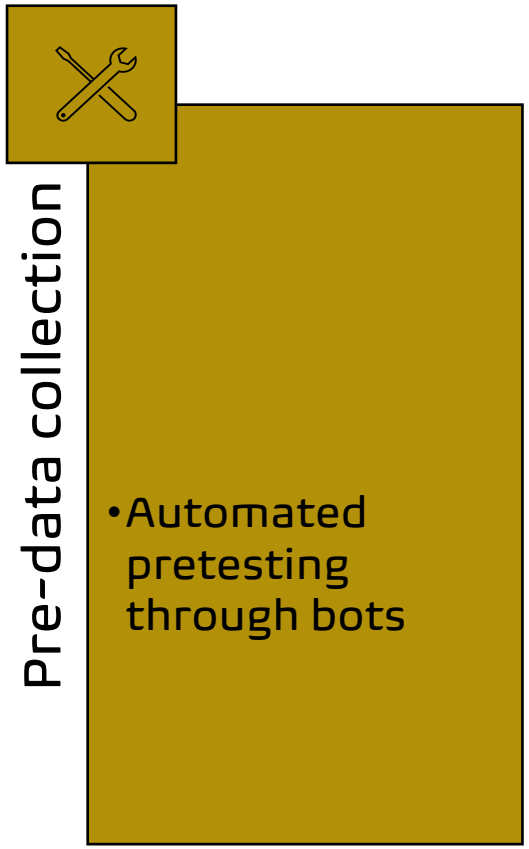
Research Areas & Future Perspectives



Research Areas & Future Perspectives



Research Areas & Future Perspectives



Höhne, J.K., Claassen, J., Shahania, S., & Broneske, D. (2025). Bots in web survey interviews: A showcase. International Journal of Market Research, 67, 3–12.

Introduction

- There is an overall trend that surveys are conducted through web modes
  - *Even large-scale survey programs, including the EUROSTUDENT, follow this trend*
  - *Web mode requires labor-intensive pretests to ensure functioning*
- Pretests require numerous trials that are usually done manually
  - *For example, surveyors repeatedly complete web surveys using predefined instructions*
  - *They test closed and open questions, attention checks, filters, randomizations, paradata etc.*
- One novel way for pretesting web surveys are bots
  - *Turning the threat of bots into a methodological merit*
  - *Bots can infinitely run through web surveys testing almost all scenarios*
  - *They can be even linked to LLMs for resembling human behavior*

Research Question (RQ)

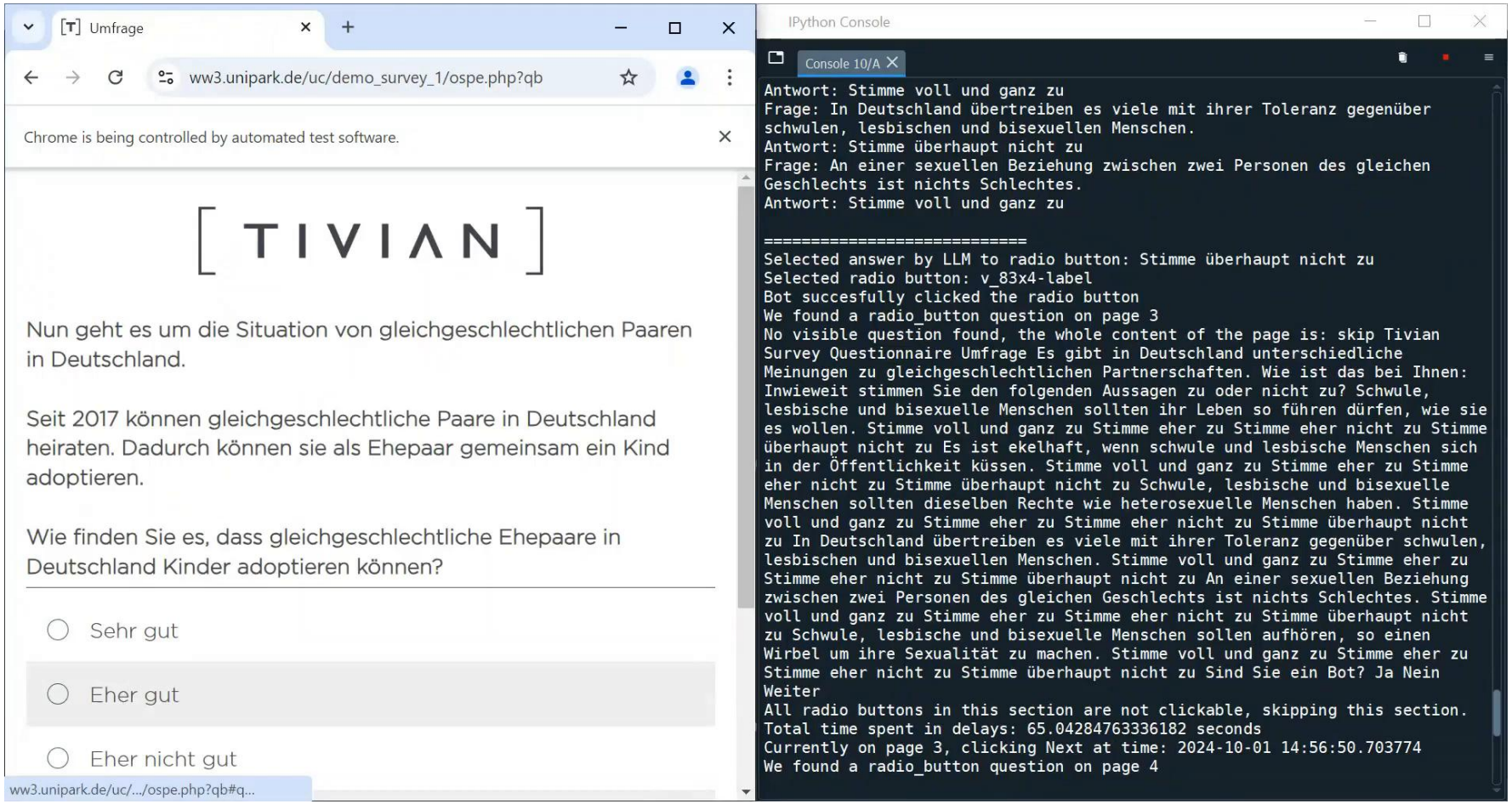
- How do bots varying in sophistication (rule- and AI-based) perform in web survey pretesting?

Bot Characteristics

- We programmed four bots with increasing capabilities
- More sophisticated bots inherit the skills of less sophisticated bots
  - *Cumulative skill sets*

Rule-based bots	AI-based bots
Rule bot + Randomly answers one question per page (per question type) + Randomly answers open text fields based on predefined strings + Tackles invisible honey pot questions	**LLM bot** (inherits Rule+ bot skills) + Classifies web survey content into opinion-based, emails, and attention checks using LLM (Gemini Pro) + Uses LLM to understand and answer questions meaningfully + Reads questions and mimics human time delay
Rule+ bot I (inherits Rule bot skills) + Handles multiple questions per page and type + Handles CAPTCHAs with text, objects, or numbers embedded in a picture + Generates random email addresses with valid domains	**LLM+ bot** (inherits LLM bot skills) + Remembers previous answers (memory) + Answers based on respondent characteristics (personas) + Handles questions with audio-visual content (speech-to-text) + Simulates paradata (mouse movements and clicks, scrolling, and keystrokes)

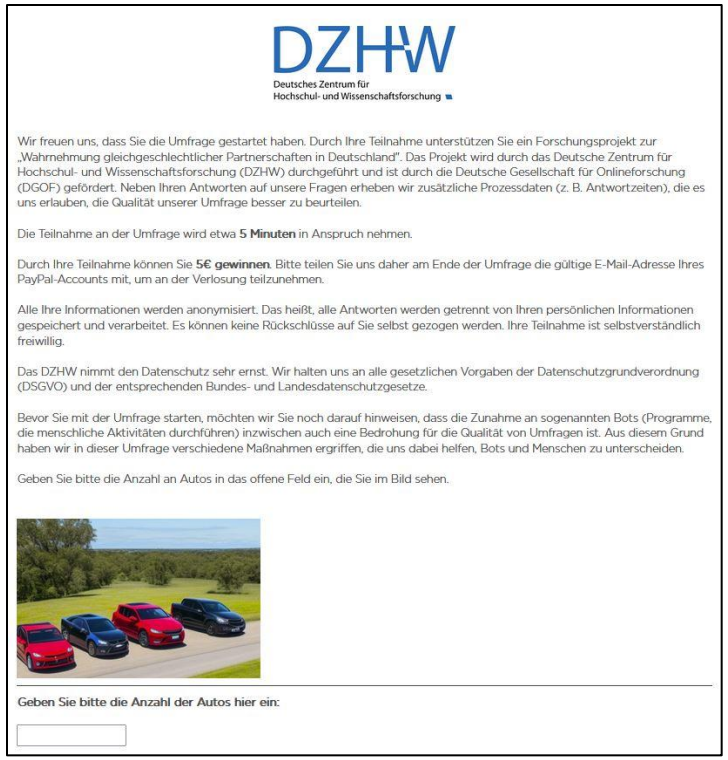
Bot Pretesting Showcase



Web Survey Design and Trials

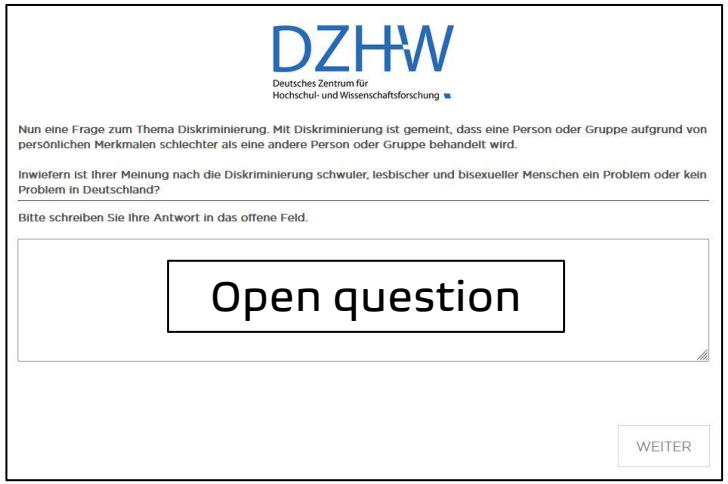
- Web survey on same-gender partnerships was programmed with Unipark
  - *Each of the four bots took the web survey 100 times (N = 400) in August 2024*
  - *Starting with the LLM+ bot, we ran the bots one-by-one through the web survey*
- The web survey included …
  - *… 3 open narrative questions*
  - *… 26 closed questions*
  - *… 1 picture CAPTCHA (counting cars)*
  - *… 2 honey pot questions*
  - *… 1 instructional manipulation check (IMC)*
  - *… 1 check-all-that-apply question (CATA)*
  - *… paradata in the form of completion times*
- The web survey included 43 questions, tasks, and instructions on 28 pages

Web Survey Screenshots I

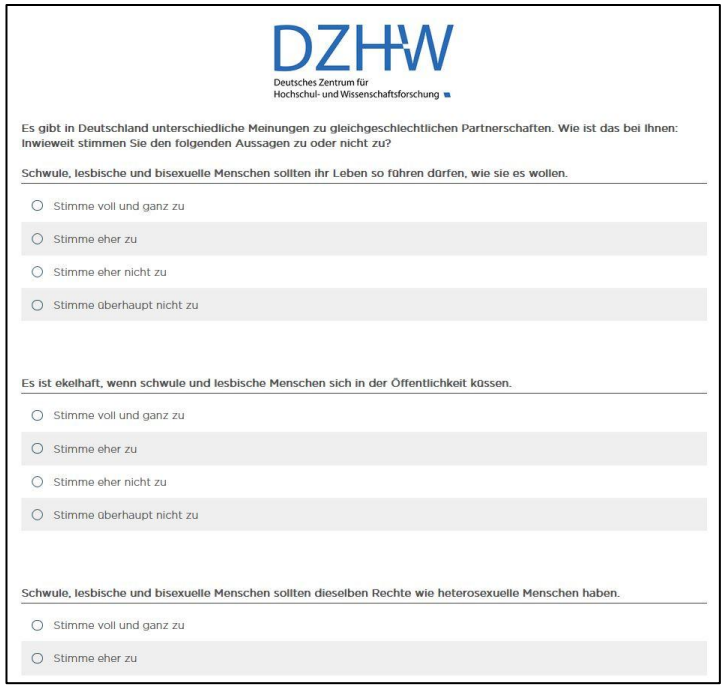


CAPTCHA

Single question

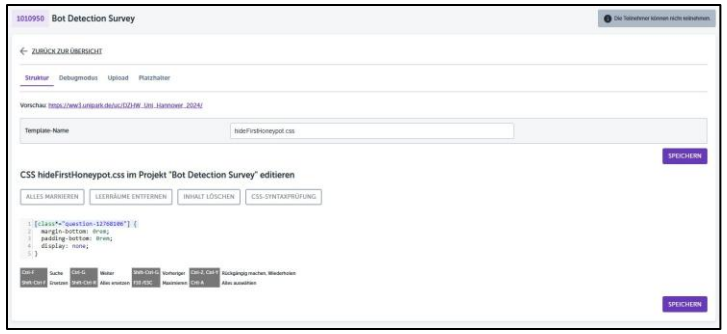


Open question

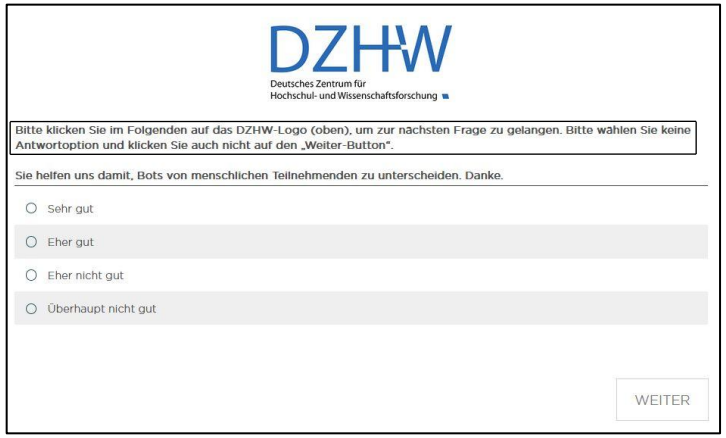


Multiple questions

Web Survey Screenshots II



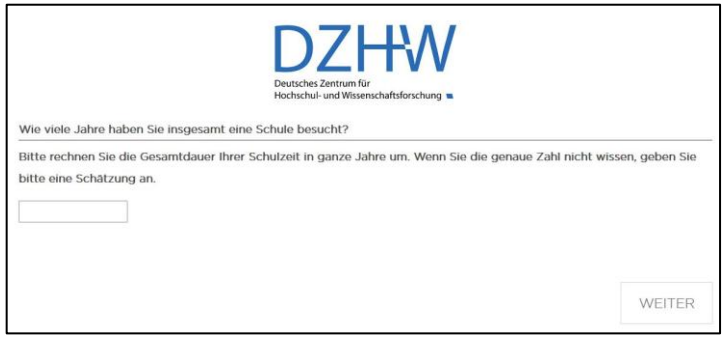
Honey pot



IMC

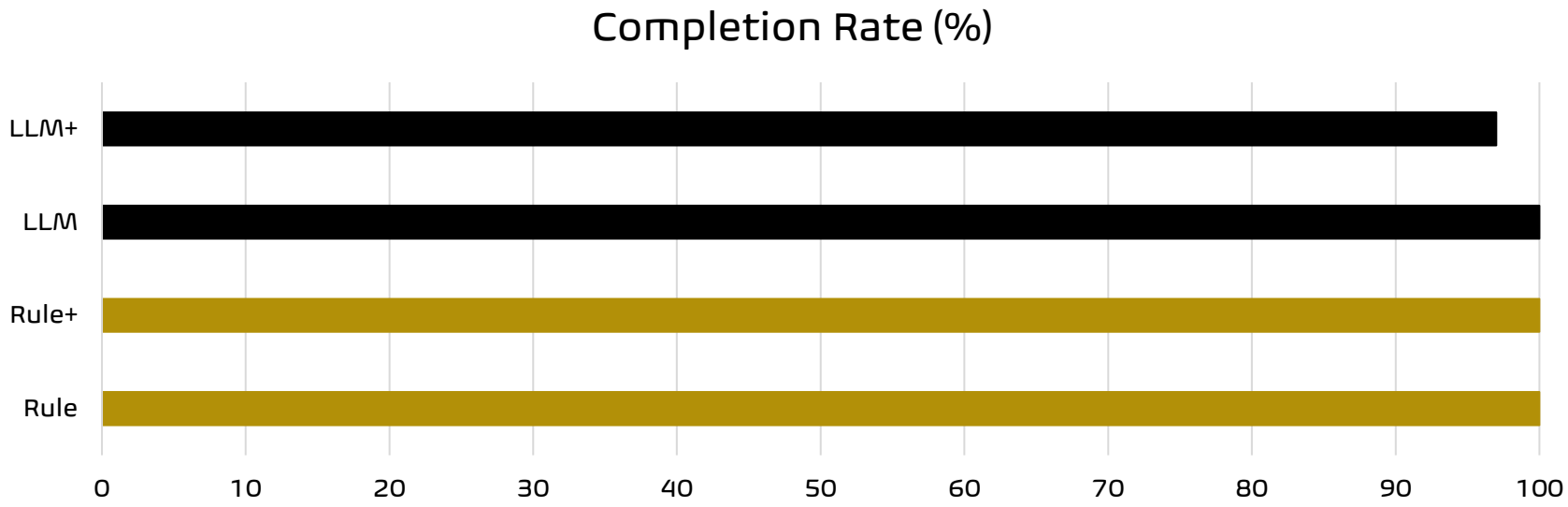
CATA

Gender



Education

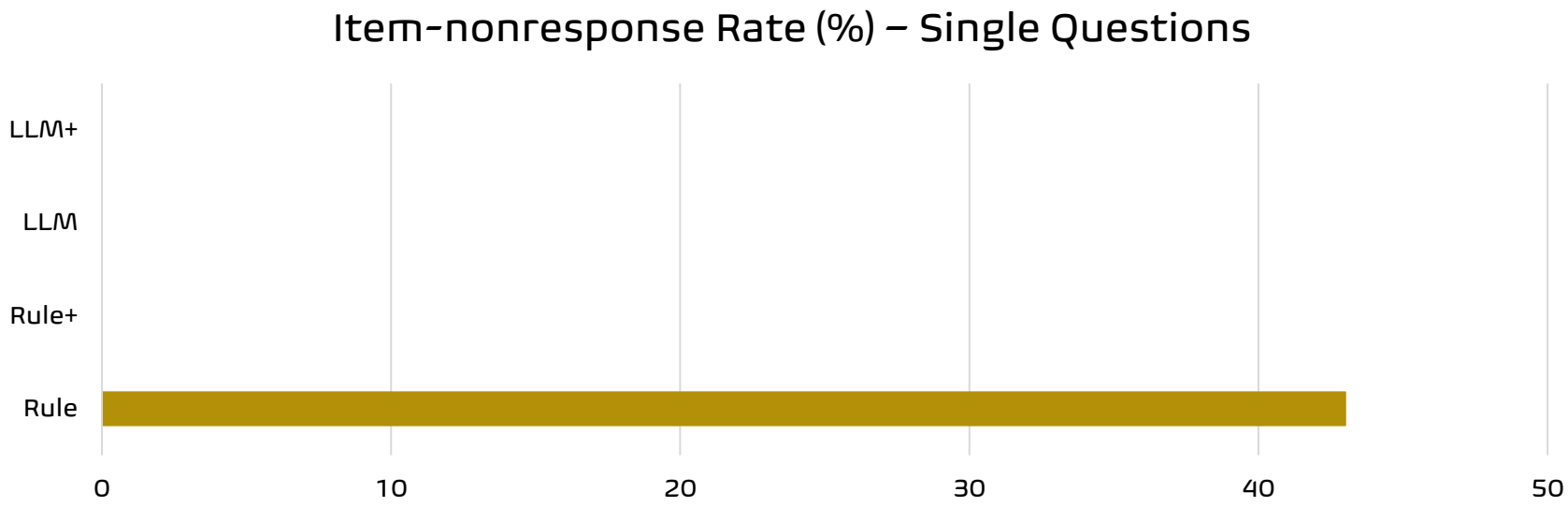
Results: Web Survey Completion



All bots complete the web survey somehow. Break-offs are very limited

Note. Rule-based bots (gold lines) and AI-based bots (black lines). Based on all 43 questions, tasks, and instructions placed on 28 web survey pages.

Results: Item-nonresponse

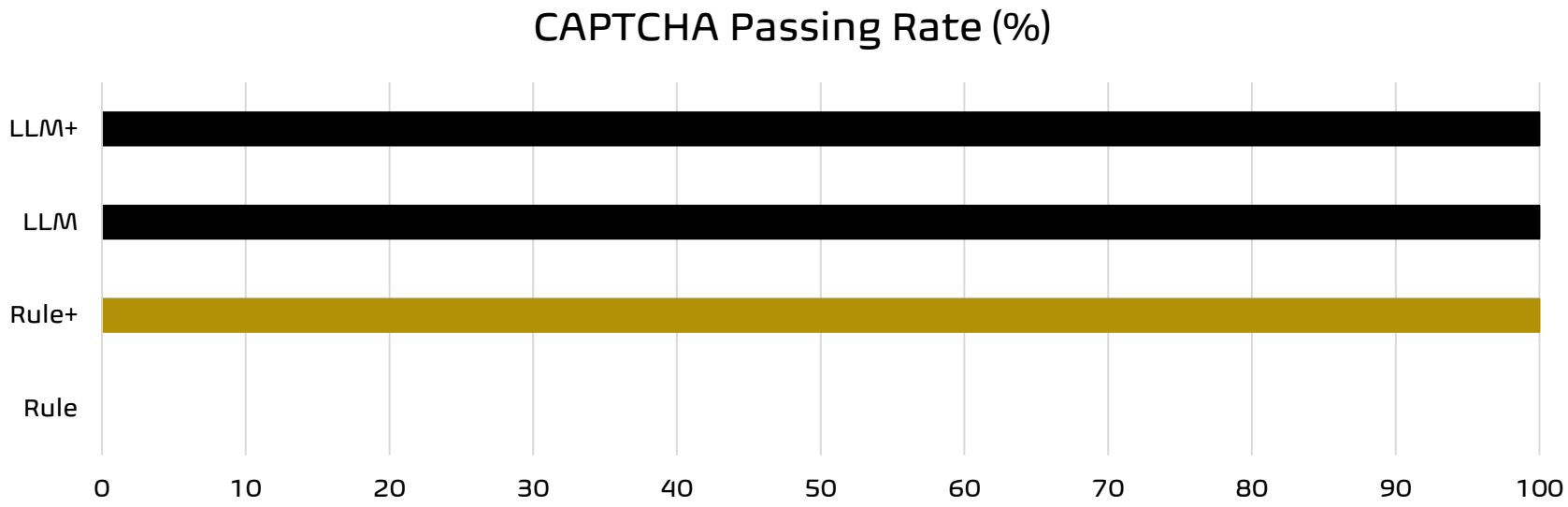


No differences between closed and open questions

Item-nonresponse is comparatively high for the basic bot

Note. Rule-based bots (gold lines) and AI-based bots (black lines). Based on 26 closed questions and three open narrative questions.

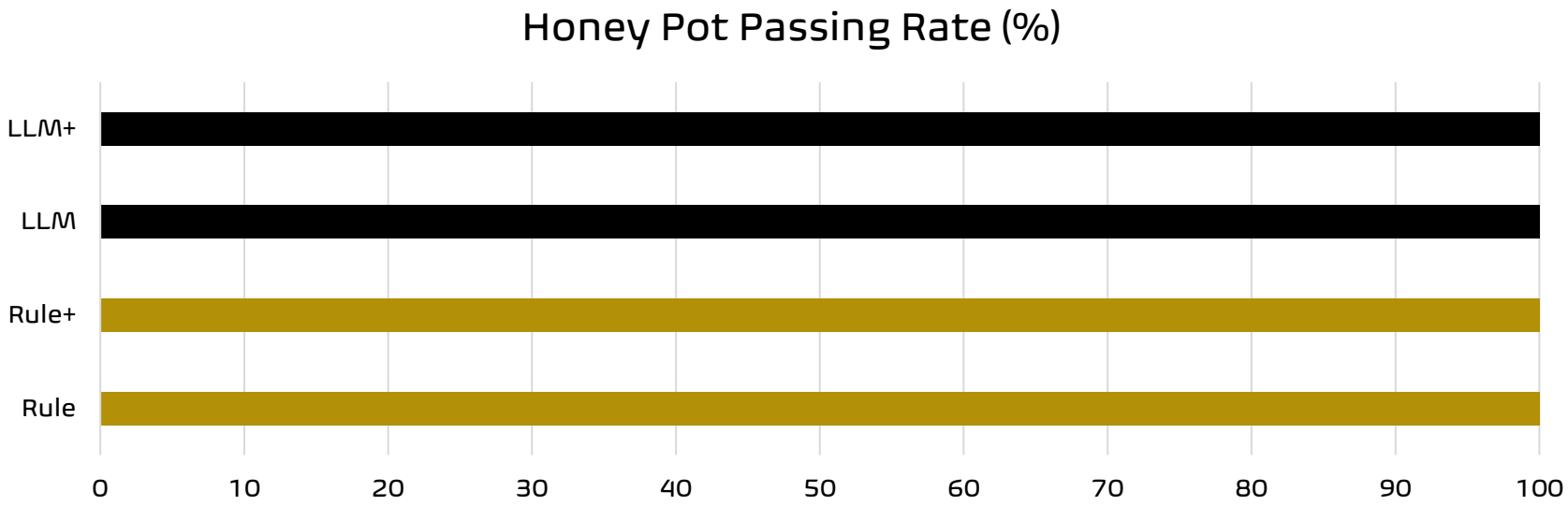
Results: CAPTCHA



Note. Rule-based bots (gold lines) and AI-based bots (black lines). Based on one CAPTCHA placed on the welcome page.

As programmed, CAPTCHAs do not constitute a problem for the bots, except for the basic one

Results: Honey Pot Questions

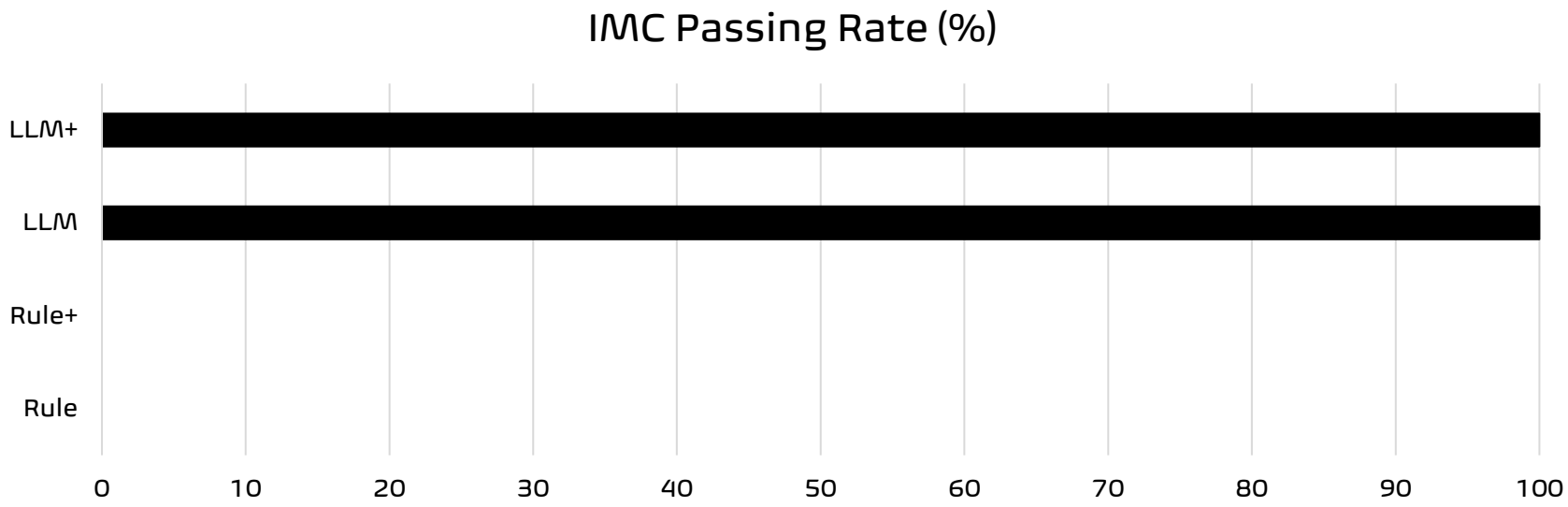


Note. Rule-based bots (gold lines) and AI-based bots (black lines). Based on two honey pot questions implemented in the source code of two web survey pages.

All bots conquer invisible honey pot questions, although they are not explicitly programmed to do so

→ Selenium WebDriver

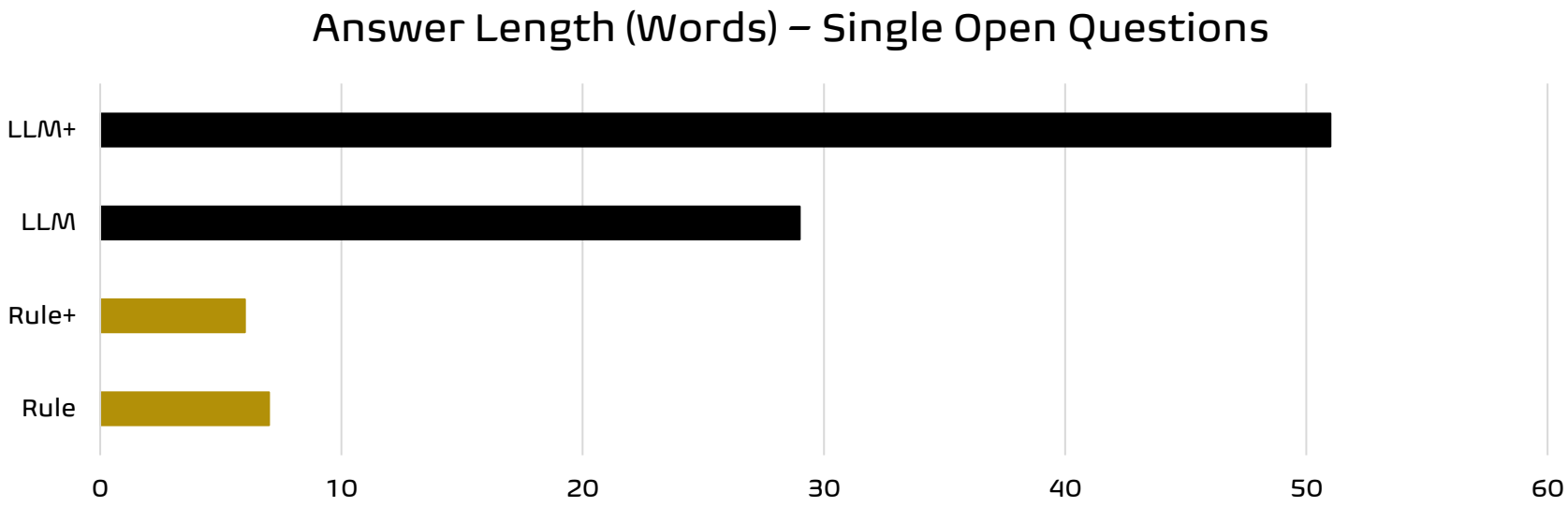
Results: IMC



IMCs only challenge the rule-based bots. The AI-based bots have understanding and thus problem-solving skills

Note. Rule-based bots (gold lines) and AI-based bots (black lines). Based on one IMC placed on one web survey page. IMC = Instructional Manipulation Check.

Results: Answer Length

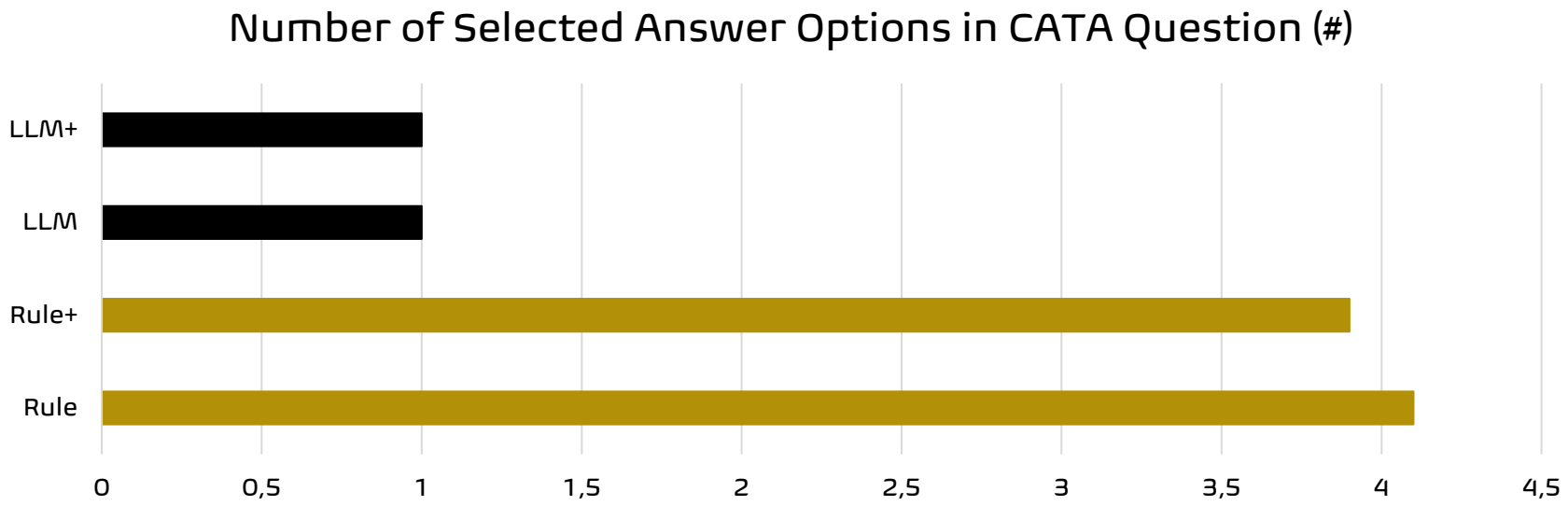


Note. Rule-based bots (Basic and Medium I) and AI-based bots (Medium II and Advanced). Based on three narrative open questions placed on three web survey pages.

Answer length increases with bot sophistication

It appears that the LLM+ bot gets "chatty"

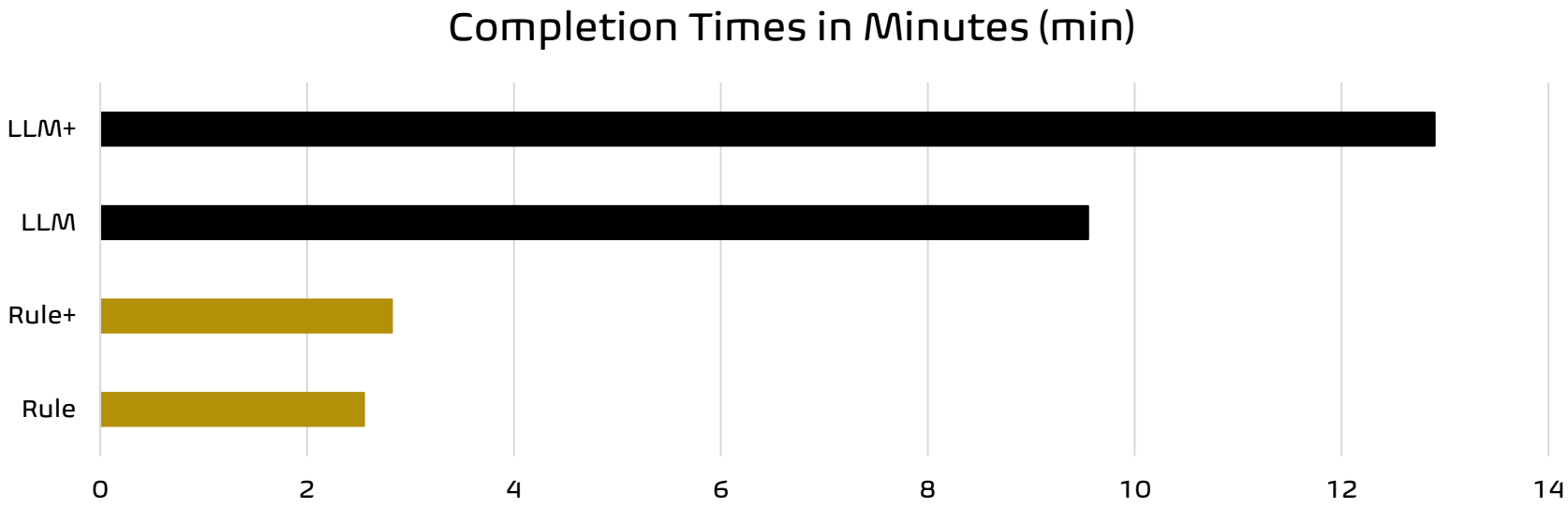
Results: CATA Question



Rule-based bots select a high number of survey locations (e.g., home, public transport, work)

Note. Rule-based bots (gold lines) and AI-based bots (black lines). Based on one CATA question on survey location placed on one web survey page. CATA = Check-All-That-Apply.

Results: Completion Times



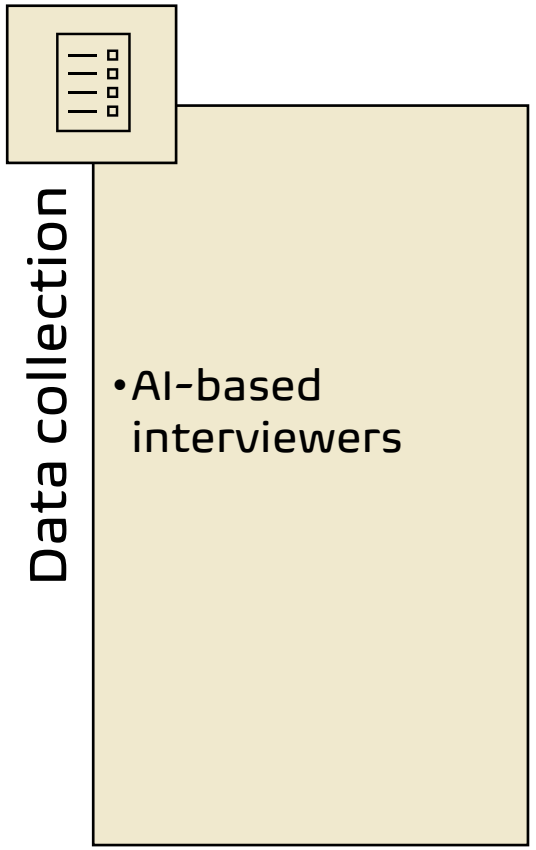
There are clear completion time differences between rule- and AI-based bots

Note. Rule-based bots (gold lines) and AI-based bots (black lines). Based on all 43 questions, tasks, and instructions placed on 28 web survey pages. We used the open-source "Embedded Client Side Paradata" tool (Schlosser & Höhne 2018).

Discussion and Conclusion

- All bots show impressive completion capabilities
  - *Completion rates are close to 100%*
- There are some clear differences between rule- and AI-based bots
  - *AI-based bots effectively test web surveys, including paradata*
  - *AI-based bots can test complex tasks, including IMC and filter questions*
  - *Rule-based bots effectively test CATA questions with multiple options*
- LLM+ bot is even capable of resembling human behavior and staying consistent
  - *Adoption of personas, such as age, gender, education, and party preference*
  - *Built-in memory feature that takes previous actions into account*
- Bots can repeat surveys an infinitely testing numerous scenarios
  - *They prove themselves highly reliable at low operation costs*

Research Areas & Future Perspectives



Höhne, J.K., Conrad, F., Neuert, C., & Claassen, J. (in progress). Exploring effects of life-like virtual interviewers on respondents' answers in a smartphone survey.

Introduction I

- Demand for high-quality data from web surveys (Knowledge Sourcing Intelligence, 2023)
- Cost-efficient and streamlined web surveys replace other survey modes, especially in-person interviews (Schober, 2018)
  - *Even large-scale social surveys start utilizing web surveys*
- Web surveys may not be good candidates for primary survey mode
  - *Depressed response rates* (Daikeler et al., 2020)
  - *Impede participation for people with low levels of literacy* (Höhne, 2023)
  - *Struggle with achieving high data quality* (Callegaro et al., 2015)
- Absence of interviewers impedes the provision of assistance and the creation of trust, motivation, and engagement

Introduction II

- Advances in communication technology and AI introduce new data collection opportunities
- Fusing elements of interviewer-based and web surveys
  - *Life-like AI-based interviewers and self-administration*
- Few studies utilized AI-based interviewers
  - *Limited knowledge about respondent satisfaction and data quality benefits*
  - *Frequently conducted in lab settings*
- Web surveys with pre-recorded interviewers have data quality benefits (West et al., 2022; Conrad et al., 2023)
  - *More disclosure of sensitive behavior, less rounding, and less error variance*
  - *Respondents connect with pre-recorded interviewers*
  - *Results may also apply to virtual interviewers*

Research Questions (RQs)

- How do AI-based interviewers affect …

  … item-nonresponse compared to text-based web surveys? ***(RQ1)***

  … answer quality compared to text-based web surveys? ***(RQ2)***

  … survey evaluations compared to text-based web surveys? ***(RQ3)***

- How are AI-based interviewers evaluated …

  … by respondents? ***(RQ4)***

AI of Interviewers

Image generation
Instruction via descriptive keywords using diffusion models or architected transformers (Zhang et al., 2023)

Text generation
Generative Pretrained Transformers and Large Language Models (Vaswani et al., 2017)

Text-to-Speech generation
Statistical Parametric Synthesis and Neural Speech Synthesis (Tan et al., 2021): **1** "text analysis" (heteronyms), **2** "voice parameter prediction" (acoustic model), and **3** "vocoder analysis" (audio snippets)

Speech and image animation
Multistep pipelines of transformers, Recurrent Neural Networks, Convolutional Neural Networks, and Generative Adversarial Networks (Chen et al., 2023)

Note. See Cheung, B. (2023): https://bennycheung.github.io/create-personal-animated-ai-avatar

Creating AI-based Interviewers

Note. Question adopted from EUROSTUDENT Round 8 and 9. See https://www.heygen.com/.

Study Design

Male casual
n = (376)

Male business casual
(n = 375)

Female casual
n = (395)

Female business casual
(n = 343)

- Experiment in a smartphone survey (N = 1,871)
- Between-subject design with 5 groups
  - *1 text control without virtual interviewer (n = 382)*
  - *4 treatment groups each with a different AI-based interviewer*
- 6 closed questions on women at the workplace (3) and family relations (3)
  - *These questions were adopted from the ESS*
- Respondents had to click ▶ for playing the AI-based interviewer video
  - *Respondents were informed that they are surveyed by AI-based interviewers*
  - *Videos could be played multiple times*

Example Question Screenshots

Starten wir mit ein paar Fragen zu Job und Karriere.

Wie finden Sie es, dass Frauen in Deutschland heutzutage immer häufiger karriereorientiert denken und handeln?

- Sehr gut
- Gut
- Weder gut noch schlecht
- Schlecht
- Sehr schlecht

Text control

- Sehr gut
- Gut
- Weder gut noch schlecht
- Schlecht
- Sehr schlecht

Male casual

- Sehr gut
- Gut
- Weder gut noch schlecht
- Schlecht
- Sehr schlecht

Female casual

Sample

Data collection was conducted in the Respondi/Bilendi panel in Germany in November and December 2023

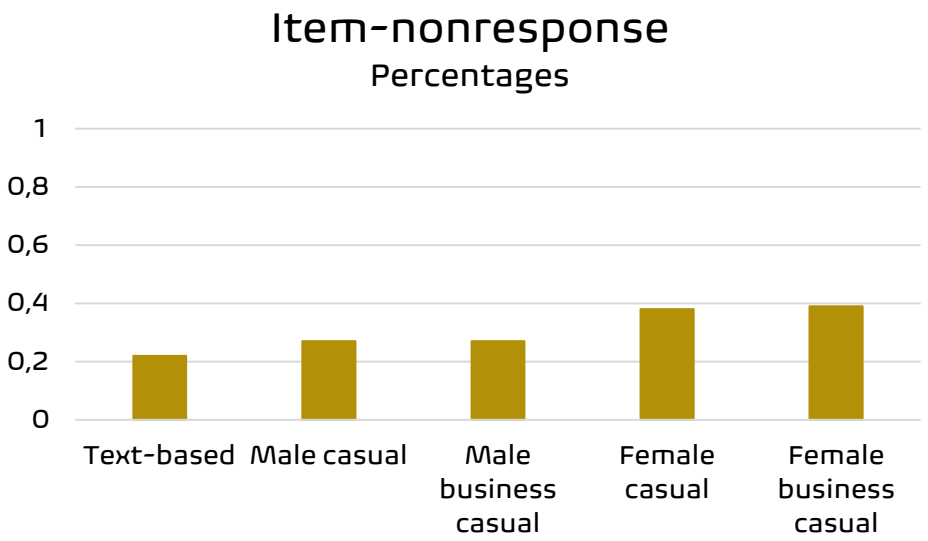
Cross quotas:	Age and gender plus quotas on education
Mean age:	49 years
Gender:	49% females
Education:	44% lower secondary school
	24% intermediate secondary school
	34% at least college preparatory secondary school

Note. Experimental groups do not statistically differ with respect to age, gender, and education

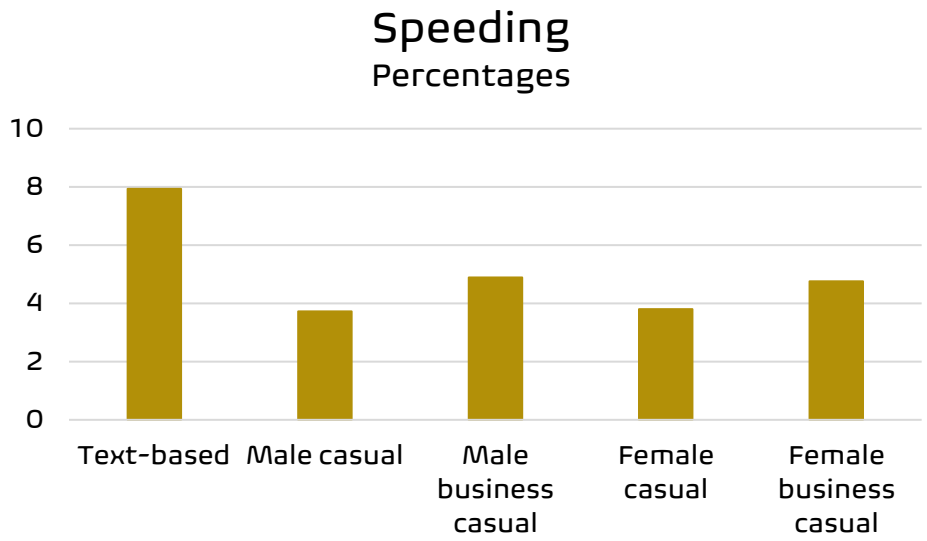
Measures

- RQ1: Determining the share of item-nonresponse – aggregated across all 6 questions
- RQ2: Determining the share of …
  … speeding (the fastest 5% percentile)
  … primacy effects (selecting the 1st option)
  … middle tendency (selecting the 2nd, 3rd, or 4th options)
- RQ3: Determining respondents' survey evaluations in terms of …
  … interest, difficulty, personal feeling, and satisfaction
- RQ4: Determining respondents' AI-based interviewer evaluations in terms of …
  … warmth, rapport, naturalness, and authenticity

Results: Research Questions 1 and 2



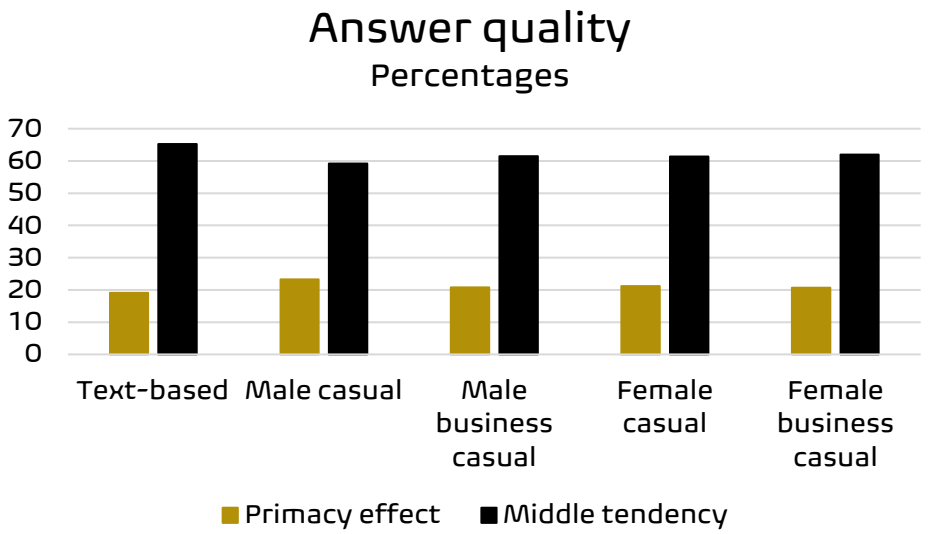
Note. We conducted *z*-tests



Note. We conducted *z*-tests

All conditions with AI-based interviewers are less prone to speeding ($p < .05$).

Results: Research Question 2

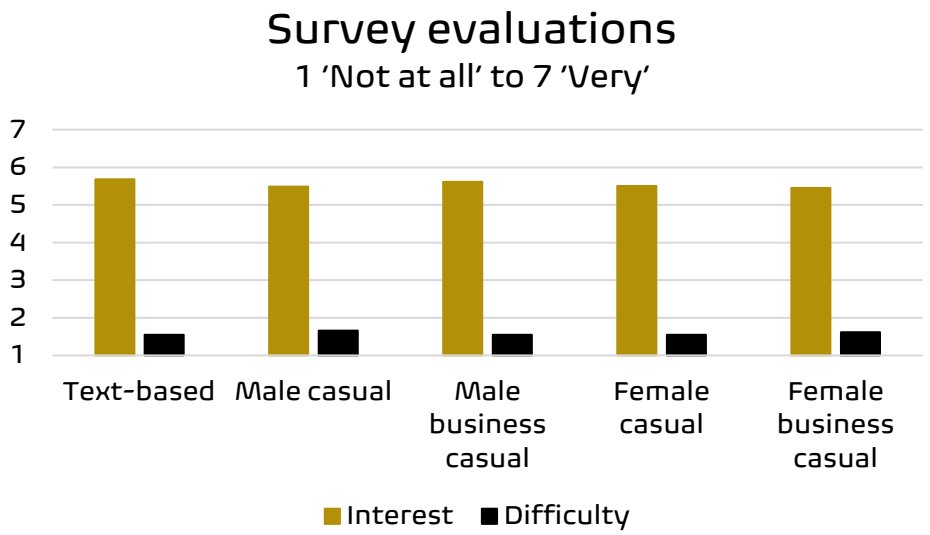


Note. We conducted z-tests

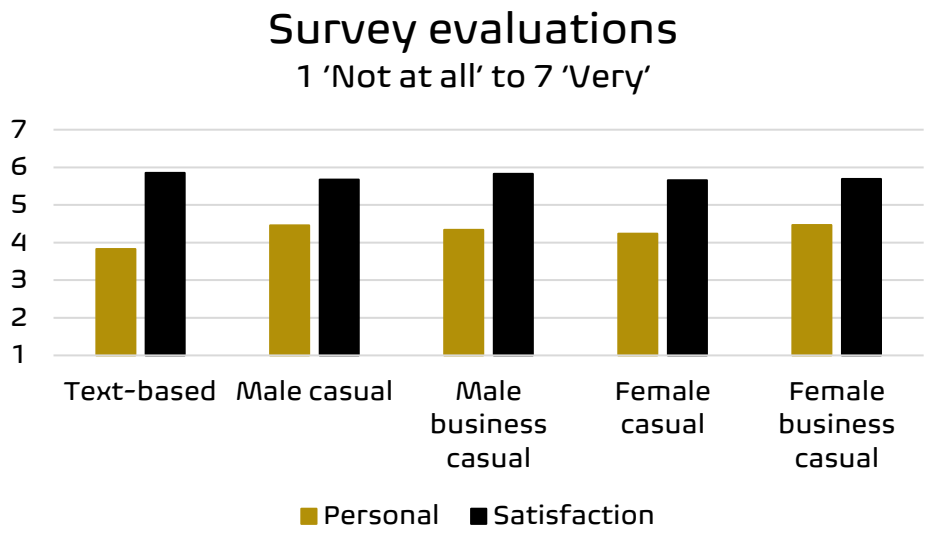
For primacy effects, we find a difference between conditions 1 and 2 ($p < .05$).

All conditions with AI-based interviewers are less prone to middle tendency ($p < .05$).

Results: Research Question 3



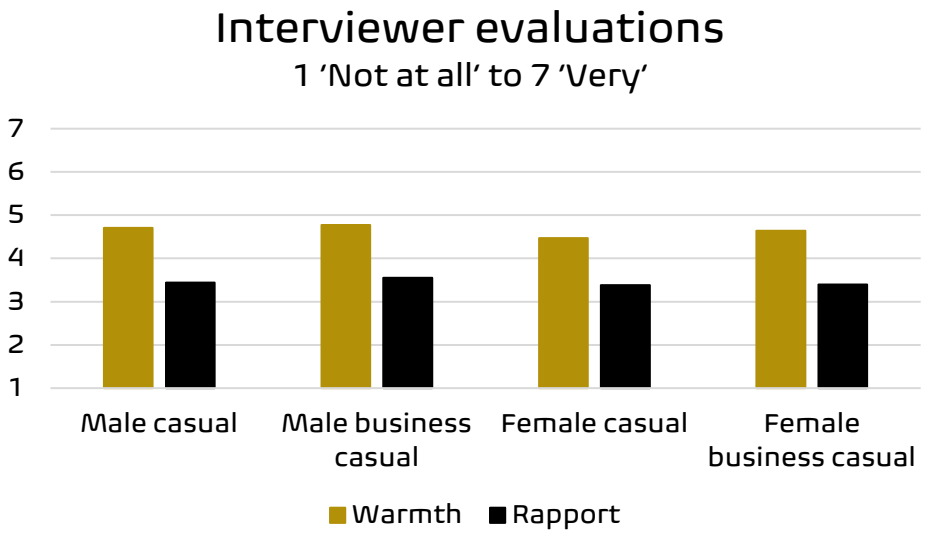
Note. We conducted ANOVAs



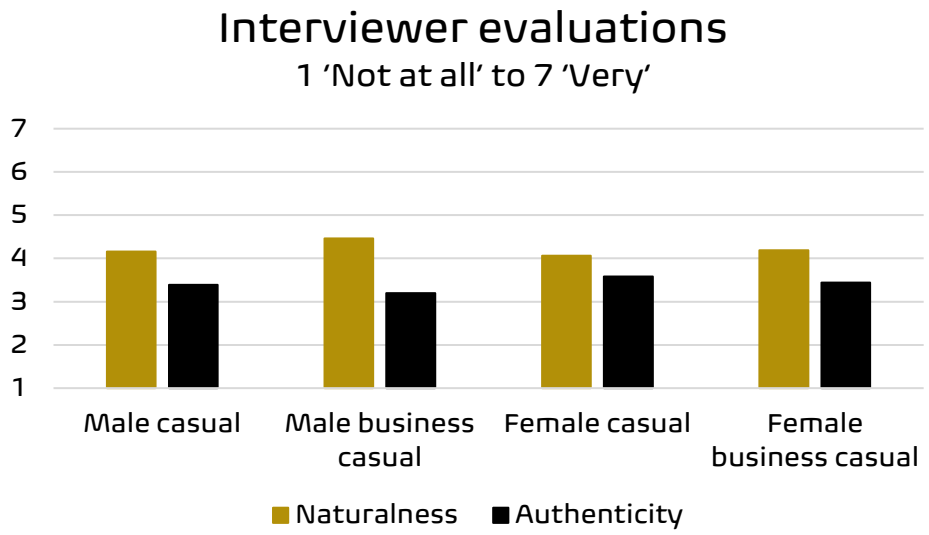
Note. We conducted ANOVAs

All conditions with AI-based interviewers are assessed as more personal ($p < .05$).

Results: Research Question 4



Note. We conducted ANOVAs

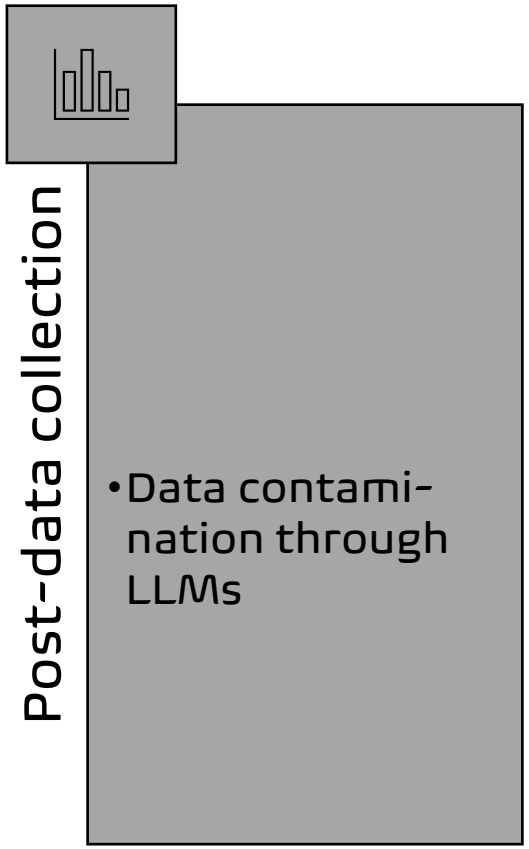


Note. We conducted ANOVAs

Discussion and Conclusion

- Item-nonresponse is low across all conditions (< 1%)
- Text-based condition is more prone to a middle tendency
  - *There are little differences regarding primacy effects*
- Speeding is much more common in the text-based condition
- All AI-based interviewer conditions are evaluated as being more personal
  - *Interest and satisfaction ratings are high (> 5 on 7-point scales)*
- AI-based interviewers perform similarly well in terms of respondent evaluations
  - *Rapport evaluations are somewhat lower than the remaining evaluations*
- Take home message: AI-based interviewers have answer quality benefits and are evaluated well by respondents

Research Areas & Future Perspectives



Claassen, J., Höhne, J.K., Bach, R., Haensch, A.C. (2026). Identifying bots through LLM-generated text in open narrative responses: A proof-of-concept study. Social Science Computer Review. Online first.

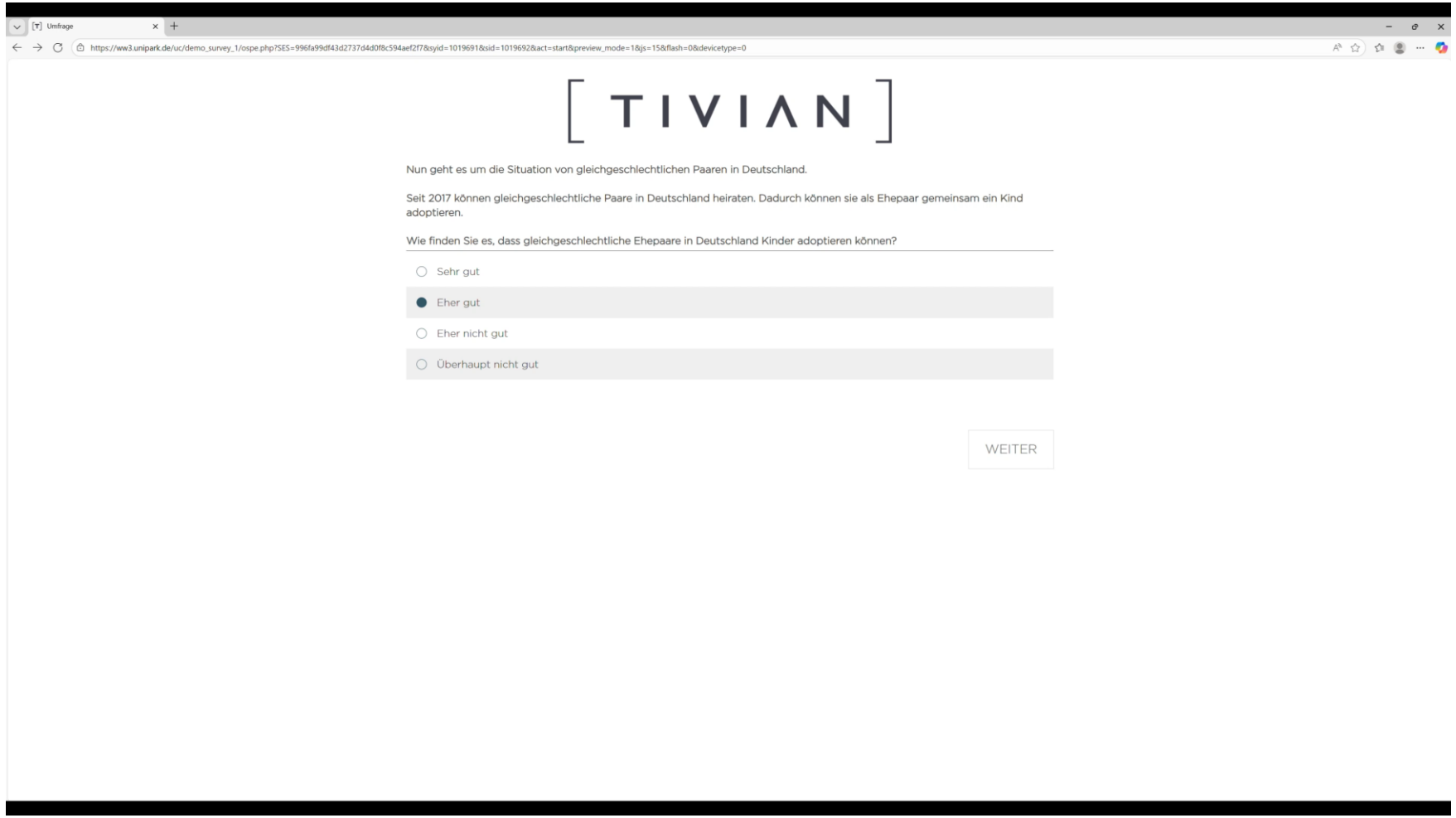
Introduction I

- Growing demand for high-quality survey data (Knowledge Sourcing Intelligence 2023)
- Cost-efficient and streamlined web surveys replace other survey modes, especially in-person interviews (Schober 2018)
- Web surveys may not be suitable for primary survey mode
  - *Depressed response rates* (Daikeler et al. 2020)
  - *Frequently struggle with achieving high data quality* (Callegaro et al. 2015)
- No interviewers for assistance and to create trust, motivation, and engagement
  - *Respondents are on their own without monitoring* (Höhne et al. 2020)
  - *Web offers numerous opportunities to cut corners: so-called "cheating"* (Scott & Jerrit 2016)
  - *The advent of Large Language Models (LLMs) has fueled the problem further* (Rilla et al. 2025)

Introduction II

- There is rumor about respondents prompting LLMs to answer open narrative questions
  - *Presenting yourself in a better light by giving elaborate answers: Social desirability*
  - *Potential threat to the quality and integrity of survey outcomes*
  - *The extent of LLM-contaminated answers and how to detect them is unclear*
- For example, the Psychological Science journal just announced that it …
  - *… "demands an explicit statement on measures taken to reduce the risk of AI-generated responses for all online studies."*
- In this study, we therefore address the following two research questions (RQs):
  - *What are the attributes of open narrative answers generated through LLMs?* ***(RQ1)***
  - *Can we detect open narrative answers in web surveys generated through LLMs?* ***(RQ2)***

Showcase: Contamination through LLMs



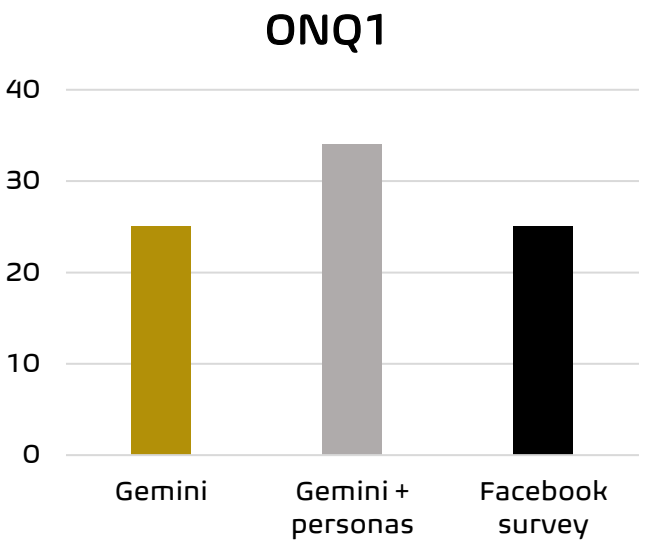
Method: Data and Analyses

- Web survey on same-gender partnerships programmed with Unipark
  - *Three open narrative questions: Child adoption, discrimination, and final comment*
  - *For each question, we prompted Gemini Pro (Google 2024) 800 times in February 2025*
  - *Gemini adopted personas – age, gender, education, and party preference – in 50% of the cases*
  - *We also conducted a web survey through Facebook (N = 1,512) in February/March 2024*
- RQ1: Text-as-data methods in the form of answer length and word choice
- RQ2: Predicting robotic language
  - *Fine-tuning BERT for each ONQ: **LLM-generated text = "yes"** or **LLM-generated text = "unclear"***
  - *Performance evaluation: Precision, recall, and F1 score*
  - *Subsequent token analysis*

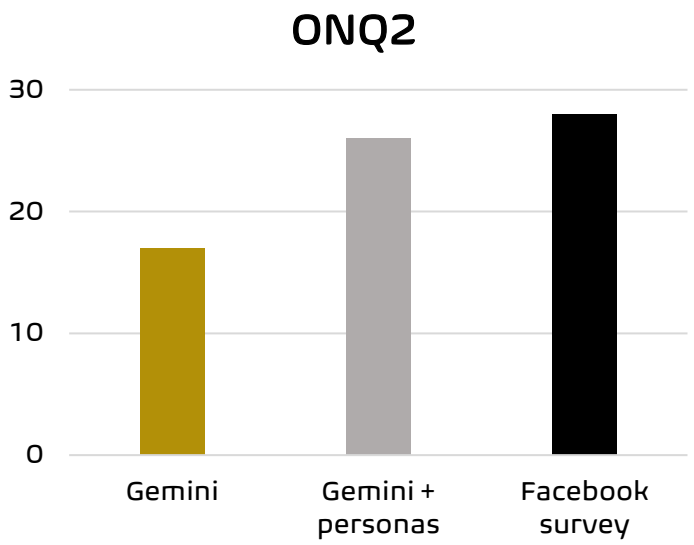
Results: Exemplary Answers

Gemini	Gemini + personas	Facebook survey
Jeder sollte die gleichen Chancen haben, eine Familie zu gründen. Liebe ist Liebe. ***Translation:*** *Everyone should have the same opportunities to start a family. Love is love.*	Ein Kind braucht 'ne Mutter und 'nen Vater. So is das nun mal vorgesehen. ***Translation:*** *A child needs a mother and a father. That's how it's meant to be.*	Hauptsache es wird sich gut um das Kind gekümmert. ***Translation:*** *The most important thing is that the child is well taken care of.*

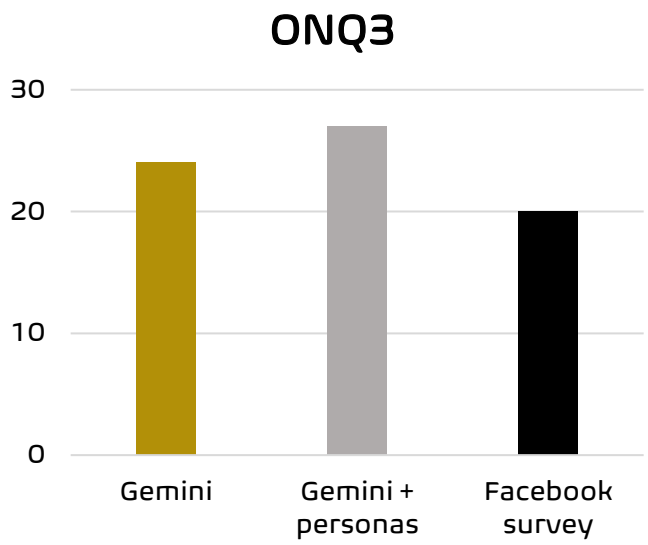
Results: Answer Length (RQ1)



Note. Average number of words.
One-way ANOVA: $p < 0.001$.



Note. Average number of words.
One-way ANOVA: $p < 0.001$.



Note. Average number of words.
One-way ANOVA: $p < 0.001$.

Results: Word Choice (RQ1)

Gemini

Gemini + personas

Facebook survey

Note. Each word cloud contains the 100 most frequently mentioned words (ONQ1) among Gemini, Gemini + personas, and the Facebook survey, respectively. The size of a word is proportional to its frequency.

Results: LLM-generated Text (RQ2)

	ONQ1	ONQ2	ONQ3
Training set size (60%)	960	960	758
Validation set size (20%)	320	320	253
Test set size (20%)	320	320	253
Precision	0.98	0.97	0.99
Recall	0.99	1.0	0.97
F1 score	0.98	0.99	0.98

Note. We used the "bert-base-german-cased" model via the "Simple Transformers" library in Python. For ONQ1 and ONQ2, we used all 800 Gemini answers as well as 800 randomly selected Facebook survey answers, respectively, to create a balanced sample. For ONQ3, in contrast, we used all 632 Facebook survey answers as well as 632 randomly selected Gemini answers.

Results: Token Analysis

	LLM-generated text = "yes"			LLM-generated text = "unclear"		
	Token	Attribution score	Frequency	Token	Attribution score	Frequency
ONQ1	(1) Fin	0.78	126	(1) auch	0.25	30
	(2) ##d	0.52	111	(2) Kinder	0.20	71
	(3) is	0.20	38	(3) Eltern	0.19	38
	(4) Ein	0.19	28	(4) und	0.17	92
	(5) ich	0.16	140	(5) zu	0.17	37
ONQ2	(1) schon	0.59	71	(1) Problem	0.31	96
	(2) Is	0.49	35	(2) nicht	0.23	73
	(3) doch	0.42	43	(3) oder	0.22	31
	(4) is	0.39	27	(4) wird	0.21	40
	(5) Also	0.39	43	(5) werden	0.20	36
ONQ3	(1) Also	0.47	46	(1) der	0.20	48
	(2) verständlich	0.43	30	(2) es	0.16	34
	(3) waren	0.27	44	(3) ##en	0.16	31
	(4) Fragen	0.25	72	(4) nicht	0.15	47
	(5) Die	0.24	39	(5) den	0.15	26

Discussion and Conclusion

- There are similarities between LLM-generated answers and those from the Facebook survey
  - *LLMs provide meaningful open narrative answers*
  - *No systematic differences regarding answer length*
  - *Word choice may offer clues when it comes to detecting LLM-generated answers*
- BERT reliably predicts LLM-generated answers
  - *Between 97 and 100 percent of the answers are correctly detected*
  - *Applies to answers from both Gemini and Gemini + personas*
- We currently explore further research possibilities
  - *Using BERT to predict prevalence of LLM-generated answers in web survey data*
  - *Making predictions based on closed questions*
  - *Examining other LLMs, such as GPT-4 and Llama 3.3*

Overall Discussion and Conclusion

- The (technological) future of student surveys remains open
- Web mode is highly amenable to technological innovations
  - *To make full use of its potential we need to form awareness and skills*
  - *Interdisciplinary cooperations (e.g., with data and computer scientists) are key*
  - *The department "Infrastructure and Methods" at DZHW is one good example*
- Importantly, this keynote only provides some potential snapshots
  - *There are further promising paths: Sensor data, trace data, data donation etc.*
  - *There are also numerous analytical innovations, including machine learning and NLP*
- Knowledge is the only resource that increases with use
  - *There are other survey programs, such as SHARE, UAS, and LISS, that we can exchange with*
  - *Others ask the same questions and face the same challenges*

Many thanks for your attention!

hoehne@dzhw.eu